

Machine Learning 1

Lecture 2.5 - Maximum A Posteriori

Erik Bekkers

(Bishop 1.2.5 - 1.2.6)



Maximum A Posteriori Estimates

- ▶ Dataset $D = (x_1, x_2, \dots, x_N)$ of N independent observations.
- ▶ ML estimate: choose \mathbf{w} such that data likelihood is maximized:

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{arg\,max}} \underset{\substack{\uparrow \\ D}}{p(\underset{\substack{\uparrow \\ \mathbf{w}}} {D} | \mathbf{w})}$$

- ▶ MAP estimate: choose most probable \mathbf{w} given the data.

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{arg\,max}} \underset{\substack{\downarrow \\ \text{posterior distribution}}}{p(\mathbf{w} | \underset{\substack{\downarrow \\ D}} {D})}$$

Curve Fitting: Maximum A Posteriori Estimates

- ▶ Dataset $D = \{(x_1, t_1), \dots, (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- ▶ Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t - y(x, \mathbf{w}))^2 \right]$
- ▶ ML estimate: choose \mathbf{w} such that data likelihood is maximized:

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta) = \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta)$$

- ▶ MAP estimate: choose most probable \mathbf{w} given the data.

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{w} | \mathbf{x}, \mathbf{t}, \beta)$$

↳ posterior

Curve Fitting: Maximum A Posteriori Estimates

▶ Dataset $D = \{(x_1, t_1), \dots, (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

▶ Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t - y(x, \mathbf{w}))^2 \right]$

▶ Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w} | \mathbf{x}, \mathbf{t}, \beta, \alpha) = \frac{p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta) \cdot p(\mathbf{w} | \alpha)}{p(\mathbf{t} | \mathbf{x}, \beta, \alpha)} \quad \leftarrow \text{does not depend on } \mathbf{w}$$

▶ Maximum A Posteriori Estimate:

$$\begin{aligned} \mathbf{w}_{MAP} &= \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{w} | \mathbf{x}, \mathbf{t}, \beta, \alpha) = \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{w} | \mathbf{x}, \mathbf{t}, \beta, \alpha) \\ &= \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta) + \log p(\mathbf{w} | \alpha) \\ &\quad \rightarrow \log p(\mathbf{t} | \mathbf{x}, \beta, \alpha) \end{aligned}$$

Curve Fitting: Maximum A Posteriori Estimates

- ▶ Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^M \mathcal{N}(w_i|0, \alpha^{-1}) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \prod_{i=1}^M e^{-\frac{\alpha}{2} w_i^2} = \left(\frac{\alpha}{2\pi}\right)^{M/2} e^{-\frac{\alpha}{2} \mathbf{w}^T \mathbf{w}}$$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} -\log p(\mathbf{w} | \mathbf{x}, \mathbf{t}, \beta, \alpha) = \underset{\mathbf{w}}{\operatorname{argmin}} -\log p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w} | \alpha)$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} -\log p(\mathbf{t} | \mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

- ▶ Curve fitting a function with Gaussian noise and Gaussian prior:

$$p(t|x, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t - y(x, \mathbf{w}))^2 \right]$$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\beta}{2} \sum_{i=1}^N (t_i - y(x_i, \mathbf{w}))^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$

- ▶ Predictive distribution:

$$p(t'|x', \mathbf{w}_{MAP}, \beta) = \mathcal{N}(t' | y(x', \mathbf{w}_{MAP}), \beta')$$

$$\mathbb{E}[t'|x', \mathbf{w}_{MAP}, \beta] = y(x', \mathbf{w}_{MAP})$$